Udacity Experiment Design – Gary Mu (12.1.15)

Metric Choice

Invariant metrics:

- **Number of cookies:** Since the cookies are randomly assigned to control and experiment group, the number of cookies in each group should be largely the same.
- **Number clicks:** The numbers of clicks on "Start Free Trial" button before the screener should be the same in control and experiment since there is no difference in user experience in the two groups at this point.
- **Click-through-probability:** Similar to the above, Click-through probability is the number of unique cookies who click on the free trial button, it is a function of #1 & #2 metric, hence this metrics should also be an invariant between the groups.

Evaluation metrics:

- **Gross conversion:** The number of user-IDs to complete checkout may decrease as the screener may have screened out users who cannot make the time commitment.
- **Retention:** The number of user-IDs remained past trial period should increase as the screener would have screened out students who cannot make the 5+ hour commitment.
- **Net Conversion:** Similar to the above, net conversion should increase if screener screens out students who cannot make the commitment.

What to expect from metrics:

As we have chosen the invariant and evaluation metrics, we would expect that invariant metrics should not change at all between groups.

For evaluation metrics, if our experiment hypothesis holds true, we should expect the number of retained users past trial period should increase, and the net conversion rate should also increase.

To summarize, the screener experiment would move the following metrics:

1. Gross conversion

H_0: $The\ screener\ will\ not\ have\ an\ effect\ on\ gross\ conversion$
H_A: $The\ screener\ will\ have\ an\ effect\ on\ gross\ conversion$

2. User retention:

H_0: $The\ screener\ will\ not\ have\ an\ effect\ on\ user\ retention$
H_A: $The\ screener\ will\ have\ an\ effect\ on\ user\ retention$

3. Net Conversion:

H_0: $The\ screener\ will\ not\ have\ an\ effect\ on\ net\ conversion$
H_A: $The\ screener\ will\ have\ an\ effect\ on\ net\ conversion$

Measuring Standard Deviation

If cookie sample size = 5000, then given the baseline conversion rate, click sample size = 400, and enrollment sample size = 82.5

- **Gross Conversion:**
  Since we are calculating a probability, we assume it follows binomial distribution, in which P = 0.20625, N = 400

  SE = sqrt(P*(1-P)/N) = 0.0202306
  The standard error of Retention Rate is 0.0202

- **Retention:**
  Since we are calculating a probability, we assume it follows binomial distribution, in which P = 0.53, N = 82.5

  SE = sqrt(P*(1-P)/N) = 0.05494901
  The standard error of Retention Rate is 0.0549

- **Net Conversion:**
  Since we are calculating a probability, we assume it follows binomial distribution, in which P = 0.1093125, N = 400

  SE = sqrt(P*(1-P)/N) = 0.01560154
  The standard error of Retention Rate is 0.0156

For Retention metric, our unit of analysis is user-id who completed checkout, and it is different from our unit of diversion (cookie), hence, the analytics variance of retention will be larger than empirical variance, hence empirical variability may be a better estimate of variance for Retention if there is time to gather this metric.

For Net Conversion and Gross conversion, since the unit of diversion and unit of analysis are the same (user cookie), hence, the analytics variance and empirical variance should be very close, and using analytical variability should be sufficient enough.

Sizing

Number of Samples vs. Power

No, I will not use the Bonferroni correction for the analysis since the metrics could be correlated with each other and move at the same time.

To determine the number of pageviews needed for the experiment, I am calculating the number of pageviews needed for each metric, and choose the largest one.

Gross Conversion
Baseline conversion: 0.20625

D_min = 0.01
Sample size = 25839
Pageview needed = (25839*40000)/3200 = 322988.5 in each group.
Total pageviews needed = 2370606*2 = 645,975

Retention:
Baseline conversion: 0.53
D_min = 0.01
Sample size = 39115
Pageview needed = (39115*40000)/660 = 2370606 in each group.
Total pageviews needed = 2370606*2 = 4,741,212

Net Conversion
Baseline conversion: 0.1093
D_min = 0.0075
Sample size = 27411
Pageview needed = (27411*40000)/3200 = 342637.5 in each group.
Total pageviews needed = 2370606*2 = 685,275

Hence, the total pageviews needed are 4,741,212

Duration vs. Exposure

The risk of this experiment is of moderate risk, because even though the screener introduces a new experience, users can still choose to start trial or just view course material after the screener. Diverting half of the daily traffic to the experiment should be Ok.

If we need 4,741,212 pageviews, and we have 40,000 pageviews per day. Diverting half of the daily traffic to the experiment would yield 238 days of experiments, and it is a long time.
Hence, we need to make new choices of only using Gross Conversion and Net Conversion as our evaluation metrics.

With these metrics, the maximum pageviews needed are 685,275. Diverting 50% of daily traffic to the experiment would require only 35 days.

Experiment Analysis

Sanity Checks

- **Number of cookies:**
  Mean proportions of cookies = 0.5
  SD proportions of cookies = 0.0006
  CI proportion of cookies = [0.4988, 0.5012]
  Observed proportions = 0.5006

  Passed sanity check

- **Number clicks:**
  Mean proportions of clicks = 0.5
  SD proportions of clicks = 0.0021
  CI proportion of clicks = [0.4959, 0.5041]
  Observed proportions = 0.5005

  Passed sanity check

- **Click-through-probability:**
  Control Click-through-probability = 0.0821
  SD proportions of cookies = 0.0005
  CI proportion of cookies = [0.0812, 0.0830]
  Experiment Click-through-probability = 0.0822

  Passed sanity check

Result Analysis

Effect Size Tests

Gross Conversion Rate:

- Pooled Gross Conversion Rate (GC): (Total Control Enrollments + Total Experiment Enrollments)/ (Total Control Clicks + Total Experiment Clicks) = 0.2086071
- Pooled GC Rate SE:
  Sqrt(Pooled GC Rate * (1- Pooled GC Rate) * (1/Total Control Clicks + 1/Total Experiment Clicks) = 0.004371675
- Difference in GC Rate:
  Diff GC Rate= Experiment GC Rate – Control GC Rate = 0.1983 – 0.2189 = -0.02055
- Diff in GC Rate 95% Confidence Interval:
  -0.02055 ± 1.96 * Pooled GC Rate SE = [-0.0291, -0.0120]

The CI does not include both 0 and practical significance (dmin = 0.01), hence the result is both statistically significant and practically significant.

Net Conversion Rate:

- Pooled Net Conversion Rate (NC):
  (Total Experiment Payments + Total Control Payments) / (Total Experiment Clicks + Total Control Clicks) = 0.2086071
- Pooled NC SE:
  Sqrt(Pooled NC Rate * (1- Pooled NC Rate) * (1/Total Control Clicks + 1/Total Experiment Clicks) = 0.004371675
- Difference in NC:
  Diff NC Rate = Experiment NC Rate – Control NC Rate= 0.1983198- 0.2188747 = -0.02055487

- Difference in NC Rate 95% Confidence Interval:
  -0.02055487 ± 1.96 * Pooled NC Rate SE = [-0.02912320 , -0.01198655]

The CI includes 0, hence the result is not statistically significant and not practically significant.

Sign Tests

Gross Conversion Rate:
Out of the 23 daily data, 19 days have Gross Conversion Rate difference smaller than 0.
Hence, the p-value for the binomial sign test is 0.002599, and the difference is statistically significantly different than 0.

Net Conversion Rate:
Out of the 23 daily data, 10 days have Gross Conversion Rate difference larger than 0.
Hence, the p-value for the binomial sign test is 0.6776, thus the difference is not statistically significantly different than 0.

Summary

For this experiment, I did not use the Bonferroni correction as the metrics (Gross conversion rate and Net Conversion Rate) could be correlated with each other, and I am using alpha = 0.05 for each metrics statistical test.

There aren't any discrepancies in my effect size and sign test, both tests showed that the experiment result for Gross Conversion rate is statistically significant, while the Net Conversion rate is not.

Recommendation

Based on the experiment result, I would not suggest to launch the experiment live, because while we see a difference in Gross conversion rate, there is no difference in Net Conversion rate.

Net conversion rate could be more important for Udacity business as it indicates if users drop out after the trial period. Since there is no movement in this metrics between control and experiment, we should not launch this experiment. Instead, we should test other experiments which could impact Net conversion rate.

Follow-Up Experiment

One thing we could not determine in the experiment is retention rate as it required a lot of pageviews, which could be be achieved in a short period amount of time. One

consideration is to lengthen the experiment until we have the required pageviews to determine if screener has effect on retention.

Besides that, the current experiment result did not see improvement in Net conversion rate. This could be due to the fact that even with the screener, users can still choose to enroll or view course material, and it turns out some students who couldn't make the commitment still enrolled, but dropped out once they started the course and "faced the music".

This could be due to students over estimated how much time they can commit vs reality.

Instead of the simple screener, we could experiement a more sophisticated screener, to inquire about their life style, work/weekend schedule, or even with some pre-requisite questions to help students understand if they can indeed make the time for the courses.